Current Opinion in **Drug Discovery & Development 2002** 5**(1):136-143**

Improving structure-linked access to publicly available chemical toxicity information

Ann M Richard[1]*, ClarLynda R Williams[1,2], and Neal F. Cariello[3]

Addresses

[1]US Environmental Protection Agency
Mail Drop 68
National Health and Environmental Effects Research Laboratories
Research Triangle Park
NC 27711
USA
Email: richard.ann@epa.gov

[2]Environmental Protection Agency Student COOP Trainee
North Carolina Central University
Durham,
NC 27707
USA

[3]Glaxo-SmithKline
PO Box 13398
Research Triangle Park
NC 27709
USA

* To whom correspondence should be addressed

Publicly available toxicity databases serve as the central resource in efforts to develop algorithms for assessing potential chemical toxicity. File standardization and linkage of chemical structures with chemical toxicity information are essential first steps in providing broad access to existing toxicity information, for deriving useful structure-activity relationship (SAR) models, performing analogue searches, and estimating the potential toxicity of new chemicals. This review will focus on current efforts to improve structure-linked access to publicly available sources of toxicity information, outlining current web-based resources as well as two new database initiatives for standardizing and consolidating public chemical toxicity information.

Keywords Toxicity databases, data sharing, structure-searchable, structure-activity relationships, SAR, DSSTox, ILSI

Introduction

Publicly available toxicity databases are invaluable community repositories of existing toxicity information, spanning diverse chemicals and toxicity endpoints of potential concern to pharmaceutical and chemical industries, and government regulatory agencies alike. They represent a significant investment of societal resources, and their contents, both in terms of chemicals and chosen endpoints, reflect the past and evolving concerns of the public, general scientific, industry and regulatory communities. A current example is the recent public and regulatory focus on endocrine disruption endpoints, leading to construction of an estrogen receptor binding database [1]. Many of these publicly available toxicity data have been generated and compiled on current chemicals in commerce by and for government agencies and, as a result, tend to be concentrated on industrial and environmental chemicals. For recent review and discussion of available toxicological information resources from US governmental and other sources, see [2**,3].

Limitations in format of current public toxicity databases

The current status of publicly available toxicity databases mirrors the broad discipline of toxicology, which spans many types of experimental investigation and information pertaining to the adverse effects of chemicals on biological systems. These databases are scattered across public and private sources, have diverse formats, and contain different types of descriptive information. Most have been compiled as resources for toxicologists and regulators and, as such, have been primarily designed as textual references on existing chemicals. An example would be the EPA Integrated Risk Information System (IRIS) database (Table 1), an important resource of toxicity summary information for use in assessing the risk of chemical exposure to humans. A major limitation of many public toxicity databases, such as IRIS, is that they do not contain chemical structure information; rather, they are most commonly indexed and searchable by common chemical names and/or CAS numbers (Chemical Abstract Service Registry numbers). CAS identifiers are non-unique, prone to transcription and formatting errors, and devoid of chemical meaning. In contrast, chemical structures have a universally understood scientific content that spans all fields of toxicological endeavor. Linkage of chemical structures with chemical toxicity information is an essential first step in deriving structure-activity relationship (SAR) models, performing analogue searches, building mechanism-based chemical groupings, and estimating the potential toxicity of new chemicals. Likewise, chemical structure indexing can impart a standardized search metric for exploring the chemical basis of toxicity within and across current toxicity databases, as well as for linking chemical toxicity information with other types of biological and chemical activity information, eg genomics data and physico-chemical properties (see eg efforts

to develop XML mark-up language [4,5]; and the sophisticated structure-search functions of the new National Cancer Institute's Structure Browser (Table 1) [6**]).

Toxicity screening and prediction

Large pharmaceutical and chemical industries have invested heavily in information technologies and data mining tools for managing, exploring, and providing widespread corporate access to large internal libraries of chemical and biological test information. Advances in object-relational data management [7], decision-support systems [8], and means for assessing database knowledge content [9], all centered on the concept of chemical structure as a key identifier and search metric, are geared towards maximizing the return of this investment. In this context, the application of information technologies to structure-based toxicity prediction is viewed as one of the most serious bottlenecks and formidable challenges to improving the success of drug candidate leads. A recent review [10*] outlines the toxicity prediction challenge from the perspective of the pharmaceutical industry and details the role that enhanced databases and computational tools are poised to make in lowering failure rates in drug development.

Unlike drug design efforts, in which a receptor-mediated event most often provides a clear focus for optimization and modeling activity, adverse effects of chemicals are generally unanticipated, can take a variety of forms, suffer from a lack of a standard lexicon, and can result from one of many possible, little-understood mechanisms. The lack of clear mechanistic definition of toxicity endpoints coupled with insufficient structural diversity of tested chemicals relative to particular toxicity mechanism categories, in turn, have been the main reasons for the limited performance and applicability of available toxicity prediction models [11-13]. Comparison of some existing toxicity prediction algorithms and commercial systems, outlining their capabilities and limitations, are provided in [14-17]. A broad survey and current description of knowledge discovery, machine learning and data mining approaches that have been applied to the analysis of toxicological data, with particular focus on chemical carcinogenicity, is provided in [18**].

Although corporate and government regulated-substances databases can be valuable sources of toxicity information on proprietary chemicals, these data have been collected for purposes other than targeted toxicity investigation and, hence, tend to be too limited for broad toxicological modeling purposes. Consolidating toxicity information on proprietary chemicals with public sources of toxicity information is clearly the ideal and will maximize the effectiveness of data mining and SAR model construction efforts. An illustrative example is provided by a toxicology/safety knowledge base and computational toxicology initiative underway within the Food and Drug Administration's Center for Drug Evaluation and Review (FDA-CDER) [19]. A structure-searchable, comprehensive inventory of toxicity information on FDA regulated substances is being created, along with the means for identifying SAR relationships. Authors have reported using confidential pharmaceutical data from FDA drug review submissions to augment public sources of toxicity data, and using these combined data to derive MultiCASE (Table 1) models for predicting rodent carcinogenicity that are more effectively tailored to FDA regulatory priorities [20*]. Analogous to the situation with corporate databases, this FDA data inventory is currently for internal FDA use only and not publicly accessible. However, a commercial version of the final FDA MultiCASE model for rodent carcinogenicity will be made publicly available. This model uses only the resulting MultiCASE "biophores" (i.e. chemical fragments) in its predictive mode, and contains no links back to the structures of the proprietary chemicals, thereby effectively shielding their identities (for additional critique of this model, see [17]). This type of model may serve as a template for future efforts to parameterize and improve commercial or public prediction models with proprietary chemical information [21,22]. In general, however, toxicity data stripped of its association with chemical structure information will have limited utility beyond its use in a fixed, validated prediction algorithm. Hence, the modeling community at-large will have to continue to rely upon, and maximize the usefulness of existing public toxicity information.

In the remainder of this review, we will focus exclusively on efforts being made to improve structure-linked access to publicly available sources of toxicity information. We will highlight some current web-based resources that are providing structure-indexed access and searching across public toxicity databases. Finally, we will outline two new database initiatives for standardizing and/or consolidating public chemical toxicity information, and for facilitating structure-based exploration of these data across the spectrum of toxicity endpoints of potential interest.

Web-based resources for structure-searchable access to public toxicity information

A full accounting of available web-based toxicity databases, their features and limitations, is outside the scope of this review. We limit our discussion to two widely used web-based public resources that are currently providing some structure-searchable access to public toxicity databases. The ChemFinder Website (Table 1) is a widely used, no-fee internet-based resource for retrieving chemical structures (2D and 3D) from chemical names and CAS numbers, that also provides structure-searchable links to hundreds of publicly available databases containing physical properties and chemical and biological activities, including a number of toxicity databases. ChemFinder maintains a large centralized list of chemical structures, associated with common names, synonyms, and CAS numbers, that it uses to cross reference to internet databases. If the user-query chemical is referenced in databases containing physico-chemical properties, these numerical properties are directly displayed in the chemical search page, along with a list of links to all member databases containing information on the chemical. The user is provided with a link to the main url of the identified database (eg the Carcinogenic Potency Database, Table 1) or, when possible, to the specific webpage containing information on the user-query chemical (eg a National Toxicology Program Technical Report, (Table 1). ChemFinder also offers some capability for generalized substructure searching to identify potential analogues to the query chemical.

Another important resource for structure-searchable access to public toxicity databases is the National Library of Medicine's (NLM's) TOXNET website (Table 1). TOXNET employs the ChemIDplus system for access to NLM structure and nomenclature

files, and allows searching by name, CAS, synonym, or structure (exact, substructure, or similarity) across NLM-member databases. Databases searchable on TOXNET include the Developmental and Reproductive Toxicology/Environmental Teratology Information Center (DART/ETIC) Database, Chemical Carcinogenesis Research Information System (CCRIS), EPA's GENE-TOX mutagenicity database, IRIS, and EPA's Toxic Chemical Release Inventory (TRI). Whereas ChemFinder provides structure-based links to outside databases, toxicity databases actually reside on the TOXNET website and, thus, are fully accessible from that site. Although some of these toxicity databases can be downloaded from NLM links through TOXNET, these exportable files do not contain chemical structures.

International Life Sciences Institute (ILSI) Toxicity SAR Database Project

The objective of the non-profit ILSI Toxicity SAR database project (Table 1) is to develop a consolidated, searchable database of toxicology testing results, spanning multiple endpoints, linked with physico-chemical data and molecular structure information. The benefits of successful completion of the project would include (i) improvement of SAR model development for predictive toxicology, (ii) use of early toxicological endpoints to predict long-term toxicological responses and (iii) data sharing and reduction of animal use in developing safety data for pharmaceutical and chemical products. We report here on preliminary progress that has been made towards this database construction.

ILSI recently established a collaboration with LHASA Limited in the United Kingdom (Table 1) to develop the Toxicity SAR database. The project is currently supported by eight private companies, primarily pharmaceutical, and three US government agencies. In addition, the British government delivered a small grant to the project. The project is to proceed through three phases: (i) a pilot phase where a database format is selected and populated with small sets of publicly-available data, (ii) evaluation of the pilot project, and (iii) proposal for development of a full-scale database. The project is in the late pilot phase, with evaluation scheduled for November of 2001. For a history of the earlier stages of the project, see [23].

Four toxicological endpoints were selected for capture in the pilot project, namely, mutagenicity, carcinogenicity, skin sensitization and hepatotoxicity. Toxicology subject matter experts from the supporting companies and agencies were charged with gathering and importing data, and advising ILSI and LHASA regarding both toxicological and database software implementation issues. The Ames *Salmonella* mutagenicity assay was selected as the first endpoint for the pilot project. At present, Ames data from the NTP database on nearly 2000 compounds have been loaded into the Toxicity SAR database [24]. Database fields exist for bacterial strains and S9 condition, making it possible, for example, to search for compounds that tested positive in *Salmonella typhimurium* TA98 in the presence of a metabolic activation system. The main source of carcinogenicity data is the Carcinogenic Potency Database (CPDB) (Table 1), containing information on over 1300 compounds [25]. Information regarding tumor type, dose levels, survival and target organs for rats and mice will be incorporated into the database, a task expected to be completed by November of 2001. The database fields for hepatotoxicity are expected to be finalized in November of 2001. It is likely that fields to capture information regarding nine liver enzymes (LDH, ALP, ALT, etc.) and two dozen pathological conditions (hypertrophy, fibrosis, cirrhosis, cholestasis, etc.) will exist. Skin sensitization data are to be derived from the Cosmetic Ingredient Review (CIR) program for approximately 800 chemicals. CIRs are published in special issues of the Journal of the American College of Toxicology and International Journal of Toxicology. The database fields for skin sensitization have not yet been defined.

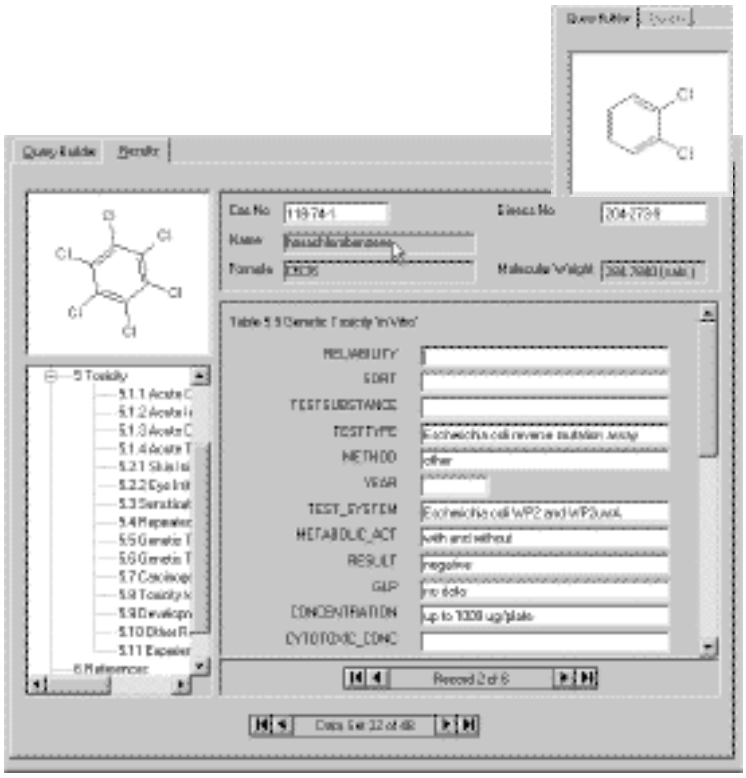
The database application format has been selected and will be a modification of the International Uniform Chemical Information Database (IUCLID) (Table 1). IUCLID is the basic tool for data collection and evaluation within the European Union - Risk Assessment Program for chemicals listed under US EPA's High Production Volume (HPV) (Table 1) volunteer testing program. In October 1999, IUCLID was accepted by the OECD (Organization for Economic Co-operation and Development) as the data exchange tool under the OECD Existing Chemicals Program. The IUCLID database system is built upon on an Oracle platform (Table 1), with an extensive data field structure. There are tables for: (i) general information, such as production site, impurities, synonyms, occupational limits and source of exposure, (ii) chemical information, such as melting point, boiling point, vapor pressure, solubility and stability in water, (iii) environmental information, such as stability in soil, biodegradation and distribution, as well as (iv) a host of toxicology endpoints such as ecotoxicity, acute and repeat dose toxicity, reproductive effects, genotoxicity, and carcinogenicity.

A major limitation of the original IUCLID database application was the lack of chemical structure fields and searchability. Chemical structures and the ACCORD search engine (Table 1) have been incorporated into the newly modified IUCLID/ILSI database that will allow for sophisticated substructure searching, including wildcard atom, list of elements at a given position, and atom features (charges, lone pair) searching. Screen shots of the search engine are provided in Figure 1. The front-end is being written in Visual Basic by software developers at LHASA. In November of 2001 a prototype is expected to be available to the companies and agencies that have been supporting the project. The prototype will be reviewed and suggestions will be made for improving the database fields, searching capabilities and the user front-end. Subsequent to these modifications, a proposal to develop the full-scale database will be circulated. Given sufficient interest and funding, efforts will proceed forward to populate the database with publicly available sources of toxicity data.

A major challenge the ILSI Toxicity SAR database effort, and any database effort that purports to try to consolidate existing toxicity information into a central database, will be to reconcile conflicting sources and interpretations of data. Either multiple reported toxicity values for the same endpoint will have to be recorded for a given chemical, or a value judgement will have to be made selecting one value over another. It is also anticipated that the final database produced in this effort will be available as a

fee-subscription service, which will be necessary to support the central management, maintenance, and continued improvements in the database.

Figure 1. Sample screen shot of the structure-searchable front end to the IUCLID Toxicity SAR database.



The small uppermost picture is the query form, where the user requested all compounds with a 1,2-dichlorobenzene substructure; no other constraints on the search were specified. The lower picture gives the results of that query in which 48 structures were found, and the 22nd structure is shown. The tree on the left-hand side has been navigated to "Genetic Toxicity in vitro", which has brought up the data in the pane on the right-hand side of the screen. Six records of "Genetic Toxicity in vitro" data for the compound have been found and the second record is shown.

Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network

Each of the above public toxicity database capabilities and initiatives suffers from some serious drawbacks from the standpoint of providing a truly public resource that adequately meets the requirements for unrestricted data access, flexible analogue searching, SAR model development, or building of a user-customized chemical relational database (CRD). The DSSTox Database Network is being proposed as a community-supported, web-based effort to attempt to meet some of these challenges. Details of the recent DSSTox proposal are provided in [26**]. For present purposes, the proposal is distilled into the following three major elements:

(i) ***Adopt and encourage the use of a common standard SDF file format for public toxicity databases.***
Structure Data File (SDF) format is a public, ASCII file format originally developed by Molecular Design Limited (currently MDL, Inc;Table 1) that stores field-delimited structure, text and property information for any number of molecules [27]. SDF was chosen for this effort because it has already been adopted as an industry-standard import/export feature of virtually all chemical modeling and CRD applications, eg ChemOffice's ChemFinder (Table 1), ACD's ChemFolder (Table 1), MDL's ISIS (Tabe 1), Accelrys's ACCORD (Table 1). DSSTox SDF file names will conform to a proposed file naming convention to communicate the origin, size, date-of-creation, and version number of the file. In addition to summary toxicity information, each DSSTox SDF file will also contain a set of standard chemical identifier fields, including structure, chemical name, CAS, SMILES, Source url, Formula Weight, Tested Form (salt, hydrate or neutral), and Substance Type (defined organic, mixture, inorganic). DSSTox SDF files are being created for a wide variety of available public toxicological databases, and will be easily convertible to data tables or importable into any commercial or private CRD application.

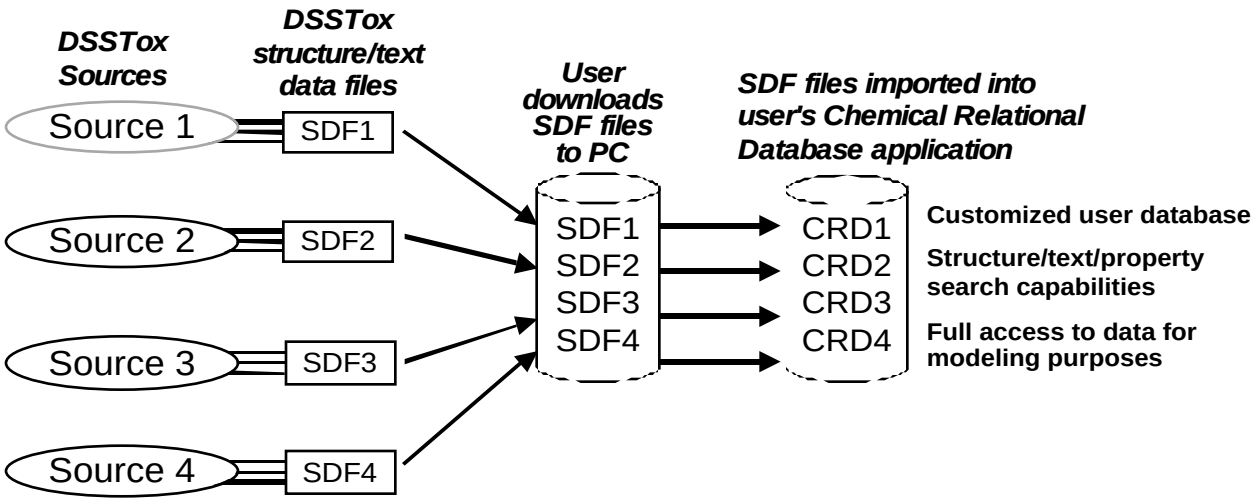
(ii) ***Implement a distributed source approach that will enable decentralized, free public access to toxicity data files, and that will effectively link toxicity data sources with potential users and modelers of these data from other disciplines.***
The DSSTox Source refers to the person(s) or organization that compiled and currently maintains a public toxicity database for which a corresponding DSSTox SDF file has been created. Ideally, the Source would be the "owner" and web-based distributor of the DSSTox SDF file, and would be asked to take responsibility for the file's maintenance and upgrade, and to document file version modifications in the DSSTox Source SDF Log file. The Source would also be referenced and acknowledged in any

subsequent use of that file. The DSSTox SDF file will provide summary toxicity information and a user will be encouraged to consult the Source website and original toxicity database for more complete textual descriptions, qualifications, references and guidance in the use of that toxicity data. Figure 2 provides a schematic illustration of how a user would retrieve DSSTox SDF files from distributed Sources to create a user-customized toxicity CRD that could be fully accessed, searched, reformulated, or merged with proprietary or other data.

(iii) ***Engage public/commercial/academic/industry groups in contributing to and expanding the DSSTox public database network.***

A DSSTox Central Website (currently under development; Table 1) will serve as the hub of the DSSTox project, providing general information on DSSTox standard file formats, a central index of field names, etc., and links to DSSTox Sources and SDF files, CRD vendors, open source scripts, and public tools and resources of general interest to the DSSTox community. Another crucial role of this website will be to connect the DSSTox user community members and to enlist their help in propagating the DSSTox recommended standards, reporting DSSTox SDF file errors to the Sources, offering enhancements to existing DSSTox SDF files, and aiding in the construction of new DSSTox SDF files.

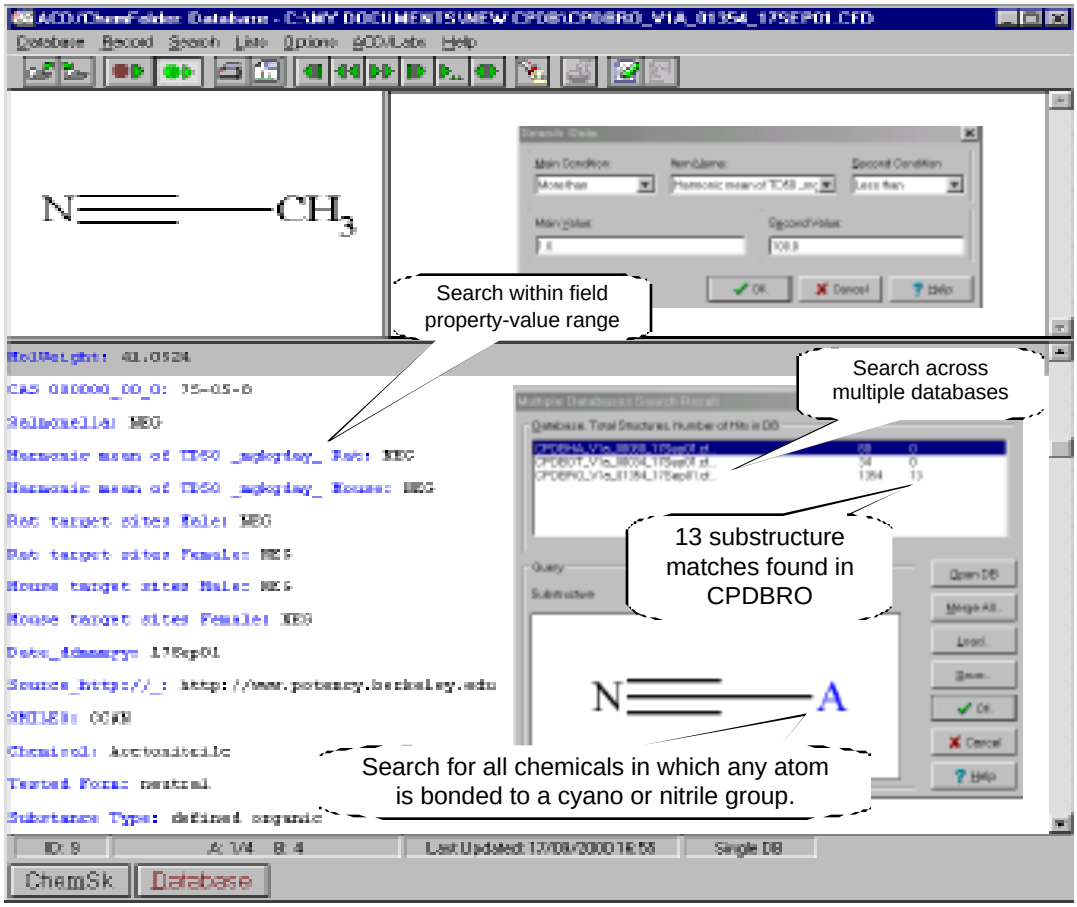
Figure 2. Steps to building a user-customized chemical relational database (CRD) from DSSTox Sources and SDF files.



DSSTox SDF files are in various stages of development for a selection of public toxicity databases spanning health and ecological endpoints. These preliminary databases were chosen for their accessibility and/or for the interest they have previously engendered from the SAR modeling community. DSSTox SDF files for the Carcinogenic Potency Database (Table 1) are near completion, and efforts have begun to create a distinct SDF file for the National Toxicology Program (NTP) rodent carcinogenicity database (Table 1). Other DSSTox SDF files currently in development pertain to the following toxicity endpoints: *Salmonella* mutagenicity (EPA/IARC Genetic Activity Profile database (Table 1) [28], EPA Gene-Tox database [29], and the NTP Salmonella database (Table 1) [24]); aquatic toxicity (EPA ECOTOX databases; Table 1); risk assessment distribution and toxicity parameters (Risk Assessment Information System; Table 1); human behavioral neurotoxicity [30]; and estrogen receptor binding [1]. Efforts are underway to complete the development of DSSTox SDF files for these databases, as well as to expand this effort outward to encompass a larger offering of DSSTox SDF files for public toxicity databases. A number of toxicity databases, eg IRIS (Table 1); and RTECS (Table 1), already exist in some form in the public domain, whereas other data sets (eg pertaining to skin sensitization, developmental toxicity, biodegradation) are available from literature or private sources and could be brought into the public domain with some community involvement and assistance. These DSSTox SDF files will be easily importable into available commercial or private CRD programs, allowing any user to create their own customized toxicity data base that can be searched across multiple fields and endpoints. A sample property search window and substructure search across multiple CPDB databases, directly imported from DSSTox SDF files into the ACD ChemFolder application, is shown in Figure 3.

The DSSTox proposal is distinguished in two important respects from those capabilities and initiatives previously discussed: 1) the complete DSSTox SDF files, including chemical structures, will exist entirely in the public domain and be freely available, allowing for completely customized use in database development; and 2) the distributed network of standard-format public toxicity databases will be a community-supported, application-independent effort, as opposed to a centralized effort creating a large application-specific database. Another clear advantage of this approach is that DSSTox Source SDF files will be faithful representations of existing databases and will not require difficult value judgements to be made on data quality or superiority of one data measurement over another. It is felt that these judgements should result from scientific discourse and consensus within the toxicological community. On the other hand, the ultimate success of the DSSTox database network initiative will depend on the active cooperation and involvement of both the toxicity database Sources and the larger DSSTox user community, both of whom stand to derive greatest benefit from its success.

Figure 3. Sample screen shot of imported DSSTox SDF database files created from the Carcinogenic Potency Database Project (Table 1), as viewed in the ACD ChemFinder application.



Search performed across multiple imported DSSTox SDF database files, as viewed in the ACD ChemFinder application; DSSTox SDF files created from the Carcinogenic Potency Database Project (Table 1).

Conclusions

Adverse effects of chemicals have the potential to span many categories and mechanisms of toxicity. Chemical structure and implied chemical reactivity characteristics, on the other hand, can serve as common conceptual framework for understanding the underlying structural basis for chemical toxicity in its many forms. Hence, imposing file standardization and improving structure-linked access to public toxicity databases are first steps to providing broad access to relevant information on a chemical's potential toxic effects. But these are, nevertheless, just first steps. It is hoped that further development of web resources and databases initiatives, such as those outlined in this review, will encourage the broader participation of chemists, computer scientists, toxicologists and others in exploring and deepening our understanding of the molecular and structural bases for toxic effects. It is only through such efforts that improved and refined models for toxicity estimation will evolve.

Disclaimer

This manuscript has been reviewed by the US Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Table 1, Toxicity database websites

Database	URL
EPA's Integrated Risk Information System (IRIS)	http://www.epa.gov/iris/
National Cancer Institute (NCI) Database Browser	http://cactus.nci.nih.gov/
MultiCASE Inc	http://www.multicase.com/
CambridgeSoft Inc. ChemFinder Website & ChemOffice ChemFinder Application	http://chemfinder.cambridgesoft.com http://products.cambridgesoft.com/family.cfm?FID=4
University of California - Berkeley, Carcinogenic Potency Project	http://potency.berkeley.edu/cpdb.html
National Institutes of Environmental Health Sciences - National Toxicology Program	http://ntp-server.niehs.nih.gov/
National Library of Medicine's TOXNET	http://toxnet.nlm.nih.gov/
International Life Sciences Institute (ILSI) & Toxicity SAR Database Project	http://www.ilsi.org http://www.ilsi.org/file/SAR.pdf
LHASA Limited, University of Leeds, UK	http://www.chem.leeds.ac.uk/luk/
European Chemicals Bureau: IUCLID database system	http://ecb.ei.jrc.it/Iuclid/
EPA's High Production Volume (HPV) Challenge Program Chemical List	http://www.epa.gov/opptintr/chemrtk/hpvchmlt.htm
Oracle Corp	http://www.oracle.com/
Accelrys Inc & ACCORD	http://www.accelrys.com/accord/ http://www.accelrys.com/offers/ci_demo/index.php
MDL Systems Inc. & MDL file formats & MDL ISIS	http://www.mdli.com http://www.mdli.com/cgi/dynamic/product.html?uid=$uid&key=$key&id=5 http://www.mdli.com/cgi/dynamic/product.html?uid=$uid&key=$key&id=30
Advanced Chemistry Development & ChemFolder CRD application	http://www.acdlabs.com http://www.acdlabs.com/download/cfolder45.html
DSSTox Central Website (currently in development, url provided for future reference)	http://www.dsstox.net
International Agency for Research on Cancer / EPA's Genetic Activity Profile Database	http://www.epa.gov/gap-db/ http://monographs.iarc.fr
EPA's ECOTOX Databases	http://www.epa.gov/ecotox/ http://www.epa.gov/med/databases/fathead_minnow.html
US Dept. of Energy, Oak Ridge National Lab - Risk Assessment Information System	http://risk.lsd.ornl.gov/rap_hp.shtml
Registry of Toxic Effects of Chemical Substances (RTECS)	http://www.cdc.gov/niosh/rtecs.html

The website urls were active and current at the time of submission of this review. If a url becomes inactive, we suggest referring to the top-level url of the company or organization to locate specific information

References

****** **of outstanding interest**
***** **of special interest**

1. Blair RM, Fang H, Branham WS, Hass BS, Dial SL, Moland CL, Tong W, Shi L, Perkins R, Sheehan DM: **The estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands**. ***Toxicol Sci*** (2000) **54**:138-153.

2. Brinkhuis RP: **Toxicology information from US government agencies.** ***Toxicology*** (2001) **157**:25-49.

** ***Thorough and current review of government sources of toxicity information resources available through the internet.***

3. Wright LL: **Searching fee and non-fee toxicology information resources: an overview of selected databases**, ***Toxicology*** (2001) **157**:89-110.

4. Murray-Rust P, Rzepa HS: **Chemical markup, XML, and the worldwide web. 1. Basic principles.** ***J Chem Inf Comput Sci*** (1999) **39**:928-942.

5. Fujita S: **XyM Markup language (XyMML) for electronic communication of chemical documents containing structural formulas and reaction schemes.** ***J. Chem. Inf. Comput. Sci.*** (1999) **39**:915-927.

6. Inlenfeldt WD, Voigt JH, Vienfait V, Oelien F, Nicklaus MC: **Enhanced CACTVS browser of the open NCI database.** ***J. Chem. Inf. Comput. Sci.*** (2001) in press.

** ***Manuscript and NCI database browser website (http://cactus.nci.nih.gov) provide example of no-fee, highly sophisticated database structure/text field searching capabilities providing valuable public access to NCI data.***

7. Cargill JF, MacCuish NE: **Object-relational databases: the next wave in pharmaceutical data management.** ***Drug Discovery Today*** (1998) **3**:547-551.

8. Ahlberg C: **Visual exploration of HTS databases: bridging the gap between chemistry and biology.** ***Drug Discovery Today*** (1999) **4**:370-376.

9. Cooke F, Schofield H: **A framework for the evaluation of chemical structure databases.** ***J Chem Inf Comput Sci*** (2001) **41**:1131-1140.

10. Johnson DE, Wolfgang GHI: **Predicting human safety: screening and computational approaches.** ***Drug Discovery Today*** (2000) **5**:445-454.
**** A good conceptual overview and discussion of need for improved toxicity prediction capabilities in drug design process.***

11. Richard AM: **Application of artificial intelligence and computational methods to predicting toxicity**, ***Knowledge Engineering Rev*** (1999) **14**:307-317.

12. Richard AM: **Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet?** ***Mutat Res*** (1998) **400**:493-507.

13. Benigni R: **The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results.** ***Mutat Res*** (1997) **387**:35-45.

14. Benigni R, Richard AM: **Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity.** ***Methods*** (1998) **14**:264-76.

15. Benfenati E, Gini G: (1997). **Computational predictive programs (expert systems) in toxicology.** ***Toxicol*** (1997) **119**:213-225.

16. Dearden JC, Barratt MD, Benigni R, Bristol DW, Combes RD, Cronin MTD, Judson PN, Payne MP, Richard AM, Tichy M, Worth AP, Yourick, JJ: **The development and validation of expert systems for predicting toxicity - the report and recommendations of ECVAM/ECB workshop 24.** ***ATLA*** (1997) **25**:223-252.

17. Richard AM, Benigni R: **AI and SAR approaches for predicting chemical carcinogenicity: Survey and status report.** ***SAR and QSAR in Environ Toxicol*** (2001) in press.

18. Helma C, Gottmann E, Kramer S: **Knowledge discovery and data mining in toxicology.** ***Statis Meth Medical Res*** (2000) **9**:1-30.
** ***An excellent overview and introduction to machine-learning and artificial intelligence approaches being applied to modeling of chemical toxicity; very thorough and balanced discussion.***

19. Matthews EJ, Benz RD, Contrera JF: **Use of toxicological information in drug design.** ***J Mol Graph Model*** (2000) **18**:605-615.

20. Matthews EJ,Contrera JF: **A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MultiCASE QSAR-ES software.** ***Regulat Pharmacol Toxicol*** (1998) **28**:242-264.
**** Details important approach and new application in predictive toxicology.***

21. Johnson DE: **The optimal fragmentation principle.** ***Drug Discovery Today*** (2001) **6**:175.

22. Richard AM: **The optimal fragmentation principle: reply**, ***Drug Discovery Today*** (2001) **6**:235-237.

23. Felton MJ: **Cooperative toxicology.** ***Modern Drug Discovery*** (Oct 2000) 81-83.

24. Zeiger E: **Carcinogenicity of mutagens: predictive capability of the Salmonella mutagenesis assay for rodent carcinogenicity.** ***Cancer Res*** (1987) **47**:1287-96.

25. Gold LS, Manley NB, Slone TH, Rohrbach L: **Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996.** ***Environ Health Perspect*** (1999) **107 Suppl 4**:527-600**.**

26. Richard AM, Williams CR: **Distributed structure-searchable toxicity (DSSTox) Public Database Network: A Proposal.** ***Mut Res*** (2002) in press.
***** First publication introducing DSSTox concept, with all details of DSSTox proposal, files, website, etc.***

27. Dalby A, Nourse JG, Hounshell WD, Gushurst A, Grier DL, Leland BA, Laufer J: **Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited.** ***J Chem Inf Comput Sci*** (1992) **32**:244-255.

28. Waters MD, Stack HF, Garrett NE, Jackson MA: **The Genetic Activity Profile database.** ***Environ Health Perspect*** (1991) **96**:41-45.

29. Auletta AE, Brown M, Wassom JS, Cimino MC: **Current status of the Gene-Tox Program.** ***Environ Health Perspect*** (1991) **96**:33-6.

30. Dick RB, Ahlers H: **Chemicals in the Workplace: Incorporating Human Neurobehavioral Testing Into the Regulatory Process.** ***Am J Ind Med*** (1998) **33**:439-453.